3.2 Logistic regression

For logistic regression analysis, evaluating four different scenarios: comparing all features with selected features and the effect of normalization and unnormalization is examined, helps to provide insights into the negative impact of redundant features on the decision-making process of the model. This approach also evaluates the effectiveness of normalization in enhancing model stability and robustness, which helps to improve prediction accuracy and generalization performance to new data. We analyzed two key performance parameters: the AUC Score and the ROC curve.

3.2.1 AUC Score

Table1 shows the AUC scores for the four scenarios:

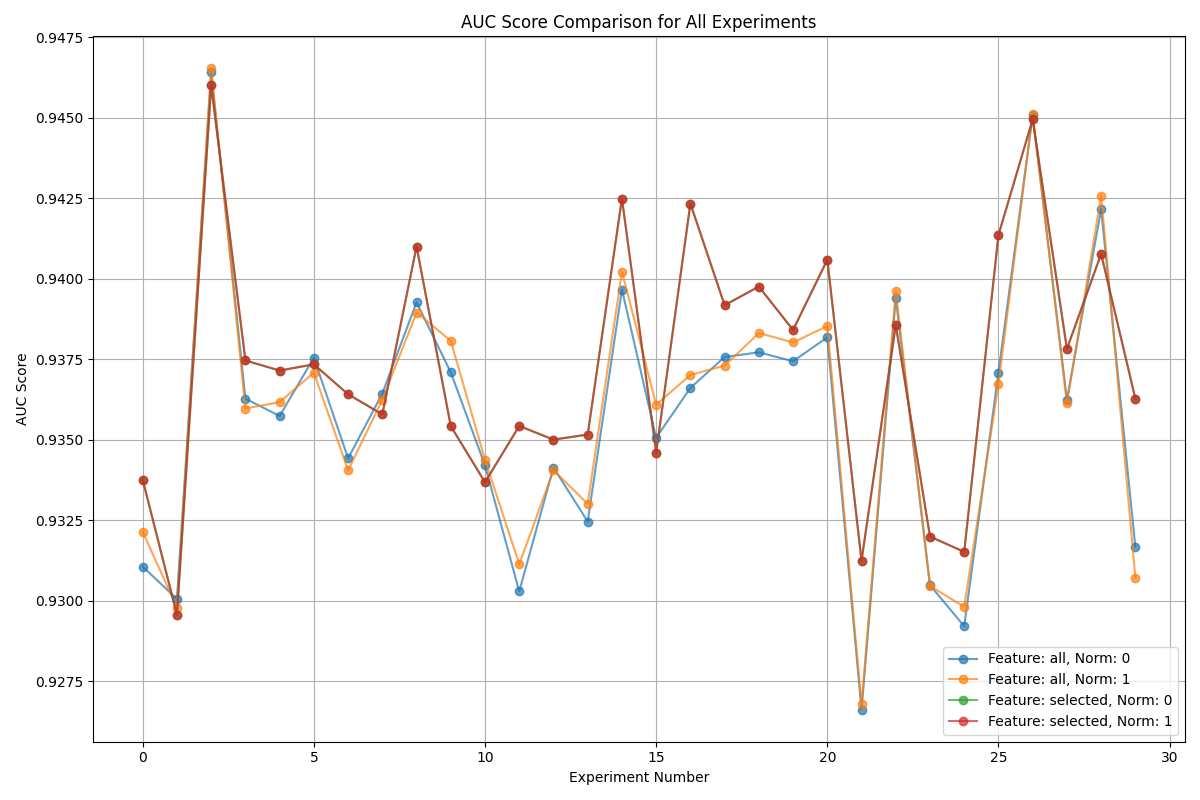
logistic\_auc\_analysis

|  |  |  |  |
| --- | --- | --- | --- |
| **feature\_type** | **norm\_flag** | **min** | **mean** |
| **all** | 0 | 0.9266109161827320 | 0.9358508302660880 |
| **all** | 1 | 0.9267741626544310 | 0.9360310771404800 |
| **selected** | 0 | 0.9295565225233350 | 0.9373666146861100 |
| **selected** | 1 | 0.9295565225233350 | 0.9373666146861100 |

Key Observations:

* All features vs. selected features: The AUC performance of selected features is slightly higher than that of all features, indicating that reducing redundant features can improve model performance in this dataset and model.
* Normalized vs. Unnormalized: According to the data, the average AUC is only about 0.0002 different between normalized and un-normalized, indicating that the standardization of features has insignificant effect on the model performance.

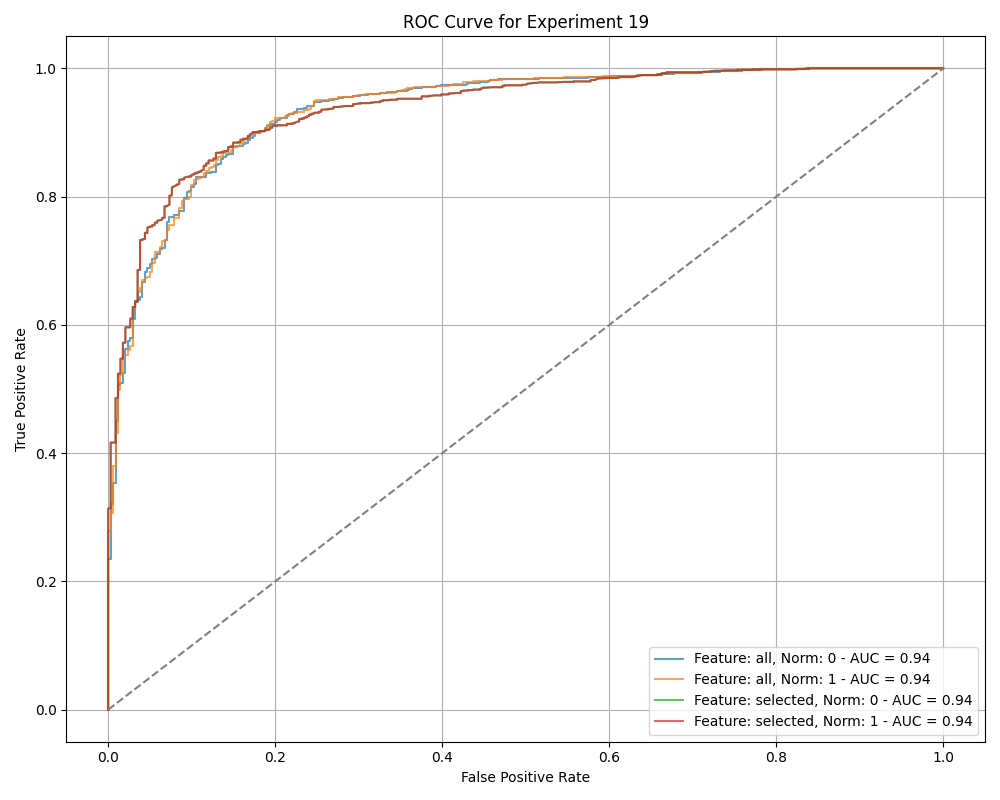
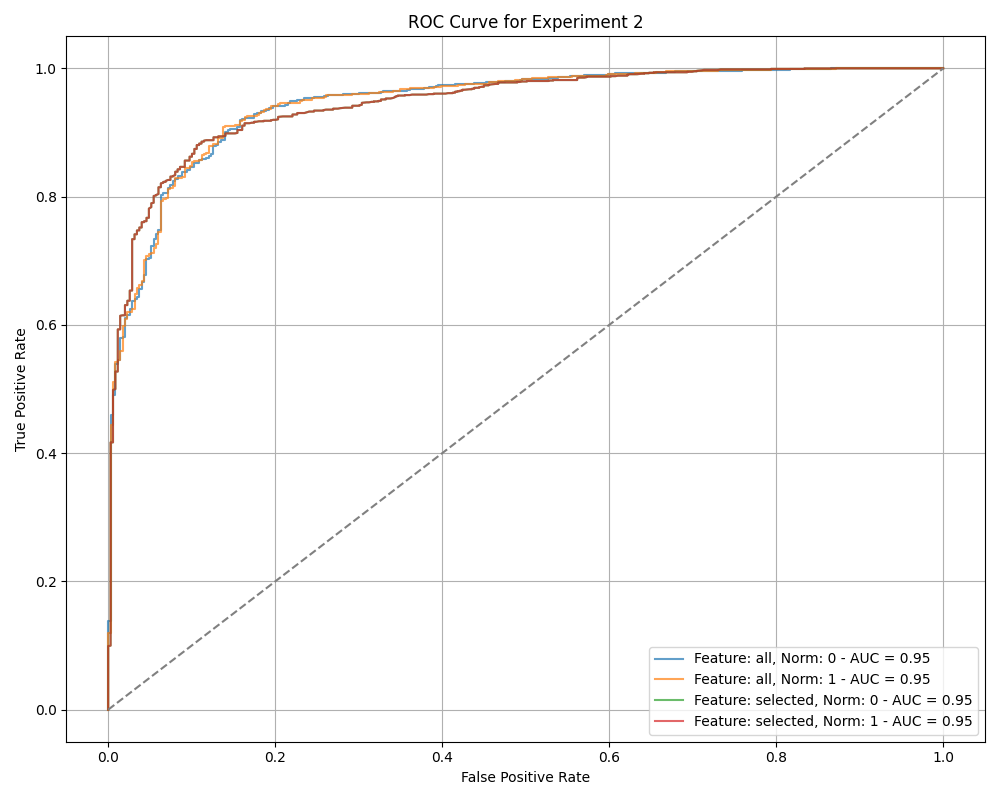
Figure1 shows the results of AUC comparison between the four contexts in 30 experiments, as shown below:



Key Observations:

The analysis results show that the AUC score of the feature selected with normalization performs more better. This may be attributed to the fact that feature selection reduces the noise interference on the model decision, while normalization ensures that the features are evaluated on the same scale, which enhances the stability and generalization ability of the model.

3.2.2 ROC curves

To further evaluate the reliability of the logistic regression model, we compared the ROC curves from 30 experimental trials. The results indicate that the area under the curve (AUC) remains consistent across trials, which suggests high robustness of the model. Below is a comparative graph of two randomly selected experiments' ROC curves: 

Key Observations:

* The ROC curves reflect the model’s performance across different thresholds, showcasing its ability to distinguish between positive and negative classes effectively.
* The consistent AUC values across experiments reinforce the model’s reliability, indicating that the logistic regression model maintains its performance regardless of variations in the training data.